# DETECTION OF OUTLIER PATCHES IN AUTOREGRESSIVE TIME SERIES

Justel, A. Peña, D and Tsay, R.S.

98-22

WORKING PAPERS

Working Paper 98-22
Statistics and Econometrics Series 15
February 1998

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

# DETECTION OF OUTLIER PATCHES IN AUTOREGRESSIVE TIME SERIES

Justel, A. Peña, D and Tsay, R.S.*

Abstract

This paper proposed a procedure to identify patches of outliers in an autoregressive process. The procedure is an improvement over the existing outlier detection methods via Gibbs sampling. It identifies the beginning and end of possible outlier patches using the existing Gibbs sampling, then carries out and adaptive procedure with block interpolation to handle patches of outliers. Empirical and simulated examples show that the proposed procedure is effective in handling masking and swamping effects caused by multiple outliers. The real example also shows that the standard Gibbs sampling to outlier detection may encounter severe masking and swamping effects in practice.

Key Words

Gibbs sampler; time series; multiple outliers; sequential learning.

*Justel, Department of Mathematics, Universidad Autónoma de Madrid; Peña, Department of Statisticis and Econometrics, Universidad Carlos III de Madrid, e-mail: dpena@est-econ.uc3m.es; Tsay, Graduate School of Business, University of Chicago.

# DETECTION OF OUTLIER PATCHES IN AUTOREGRESSIVE TIME SERIES

Ana Justel* Daniel Peña** and Ruey S. Tsay***

*Department of Mathematics, Universidad Autónoma de Madrid

**Department of Statistics and Econometrics, Universidad Carlos III de Madrid

***Graduate School of Business, University of Chicago

**Abstract**

This paper proposed a procedure to identify patches of outliers in an autoregressive process. The procedure is an improvement over the existing outlier-detection methods via Gibbs sampling. It identifies the beginning and end of possible outlier patches using the existing Gibbs sampling, then carries out an adaptive procedure with block interpolation to handle patches of outliers. Empirical and simulated examples show that the proposed procedure is effective in handling masking and swamping effects caused by multiple outliers. The real example also shows that the standard Gibbs sampling to outlier detection may encounter severe masking and swamping effects in practice.

*Key words:* Gibbs sampler. Time series. Multiple outliers. Sequential learning.

# 1. INTRODUCTION

Outliers in time series can have adversed effects on model identification and parameter estimation. Fox (1972) defined two different types of outliers in a univariate time series, namely the innovational and additive outliers. Chang and Tiao (1983) show that additive outliers can produce serious biases in parameter estimation whereas innovative outliers only have minor effects in estimation. In this paper we deal with additive outliers when they occur in patches in an autoregressive (AR) process. The main motivation of our study is that multiple outliers often have severe masking and swamping effects that may render the usual outlier detection methods ineffective.

Several procedures are available in the literature to handle outliers in a time series. Chang and Tiao (1983), Chang, Tiao and Chen (1988) and Tsay (1986, 1988) proposed an iterative procedure to detect four types of disturbance in an ARIMA model; Denby and Martin (1979), Martin, Samarov and Vandaele (1983) and Bustos and Yohai (1986) studied robust estimation; Peña (1987, 1990) proposed diagnostic statistics to measure the influence of an observation, and McCulloch and Tsay (1994) used Gibbs sampling to detect outliers and to estimate parameters of an AR process. However, these procedures may fail to identify multiple outliers due to masking effects. They can also misspecify "good" data points as outliers. This latter erroneous inference is commonly referred to as the swamping or smearing effect. Furthermore, similar to the case of independent data, influence measures based on data deletion (or equivalently, using technique of missing values in time series analysis) will encounter difficulty due to masking. Chen and Liu (1993) proposed a modified iterative procedure to reduce masking effects by estimating jointly the model parameters and the magnitudes of outlier effects. This modified procedure may also fail since it starts with parameter estimation that assumes no outliers in the data, see Sánchez and Peña (1997). Bruce and Martin (1989) were the first to analyze patches of outlier in a time series. They proposed a procedure to identify outlying patches by deleting blocks of consecutive observations. However efficient procedures to determine the block sizes and to carry out the necessary computation have not been developed.

McCulloch and Tsay (1994) showed that Gibbs sampling provides accurate parameter estimation and effective outlier detection for an AR process when the additive outliers are not in patches. However, as clearly shown by the following example, the usual Gibbs algorithm may fail when the outliers occur in a patch. Consider the outlier-contaminated time series shown in Figure 1. The outlier-free data consist of a random realization of $n = 50$ observations given in Table 1 and generated from the AR(3) model,

$$x_t = \begin{cases} a_t & t = 1, 2, 3 \\ 2.1x_{t-1} - 1.46x_{t-2} + 0.336x_{t-3} + a_t & t = 4, \ldots, 50, \end{cases}$$

where $\{a_t\}$ is a sequence of independent and identically distributed Gaussian variates with mean zero and variance $\sigma_a^2 = 1$. The roots of the autoregressive polynomial are 0.6, 0.7 and 0.8 so that the series is stationary. A single additive outlier of size 7 has been added to the time index $t = 27$, and a patch of four consecutive additive outliers of sizes 20, 20, 17 and 15, respectively, have been introduced at the indexes from $t = 38$ to $t = 41$. Assuming that the AR order $p = 3$ is known, we performed the usual Gibbs sampling to estimate model parameters and to detect outliers. Figure 2 gives some summary statistics of Gibbs sampling using the last 1,000 iterations of a Gibbs sampler with 31,984 iterations. Figure 2(a) shows the posterior probabilities of being an outlier for each data point, and Figure 2(b) gives the posterior means of outlier sizes. From the plots, it is clear that the usual Gibbs sampler easily detects the isolated outlier at $t = 27$ with posterior probability of an outlier close to one and posterior mean of outlier size 7.09. On the other hand, the Gibbs sampler fares poorly in detecting the patch of additive outliers. Specifically, the usual Gibbs sampler encounters several difficulties. First, it fails to detect the inner points of the outlying patch as outliers; the outlying posterior probabilities are very low at $t = 39$ and 40. This phenomenon is referred to as masking effects. Secondly, the sampler misspecifies the "good" data points at $t = 37$ and 42 as outliers because the outlying posterior probabilities of these two points are close to unity. The posterior means of the sizes of these two erroneous outliers are $-5.24$ and $-2.71$, respectively, for $t = 37$ and 42. In

| $t$ | $x_t$ | $t$ | $x_t$ | $t$ | $x_t$ | $t$ | $x_t$ | $t$ | $x_t$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.1274 | 11 | -14.0664 | 21 | 4.3355 | 31 | 1.4819 | 41 | -4.9735 |
| 2 | 0.5541 | 12 | -14.4163 | 22 | 5.0138 | 32 | -3.4934 | 42 | -7.3584 |
| 3 | -1.0973 | 13 | -12.9740 | 23 | 5.2705 | 33 | -8.6292 | 43 | -9.7666 |
| 4 | -3.8683 | 14 | -9.8910 | 24 | 5.6541 | 34 | -10.6260 | 44 | -9.4669 |
| 5 | -4.9750 | 15 | -6.6001 | 25 | 5.5110 | 35 | -10.1068 | 45 | -7.8670 |
| 6 | -5.8816 | 16 | -3.3033 | 26 | 5.4403 | 36 | -8.0770 | 46 | -5.2135 |
| 7 | -6.2186 | 17 | -1.6699 | 27 | 7.8908 | 37 | -5.4360 | 47 | -2.8650 |
| 8 | -7.8022 | 18 | -1.3128 | 28 | 10.0692 | 38 | -5.9219 | 48 | -0.0087 |
| 9 | -9.7464 | 19 | -0.1391 | 29 | 9.9644 | 39 | -5.5209 | 49 | 0.9087 |
| 10 | -12.0397 | 20 | 1.9578 | 30 | 6.3423 | 40 | -4.8099 | 50 | 1.9833 |

**Table 1**: Artificial time series data without outliers.

short, two "good" data points at $t$ = 37 and 42 are swamped by the patch of outliers. Thirdly, the sampler correctly identifies the boundary points of the outlier patch at $t$ = 38 and 41 as outliers, but it substantially underestimates their sizes; the posterior means of the outlier sizes are only 5.11 and 5.61, respectively, for $t = 38$ and 41. Finally, the Gibbs sampler provides biased parameter estimates. Based on the posterior means of the last 1,000 Gibbs iterations, the estimated model is

$$y_t = -0.09 + 2.16y_{t-1} - 1.82y_{t-2} + 0.58y_{t-3} + e_t$$

with estimated innovational variance $\hat{\sigma}_e^2 = 2.15$.

The above simple example clearly illustrates the masking and swamping problems encountered by the usual Gibbs sampler when additive outliers exist in patches. Similar problems also occur in the linear regression case; see Justel and Peña (1996). Consequently, further research is needed in order to effectively handle patches of outliers in a time series. The objective of this paper is to propose a new procedure that can detect the locations and sizes of patches of additive outliers and, hence, provide posterior estimates of the parameters that are free of outlier effects. Our approach is an improvement over the usual Gibbs sampler. Limited experience shows that the proposed approach is effective in handling patches of outliers.

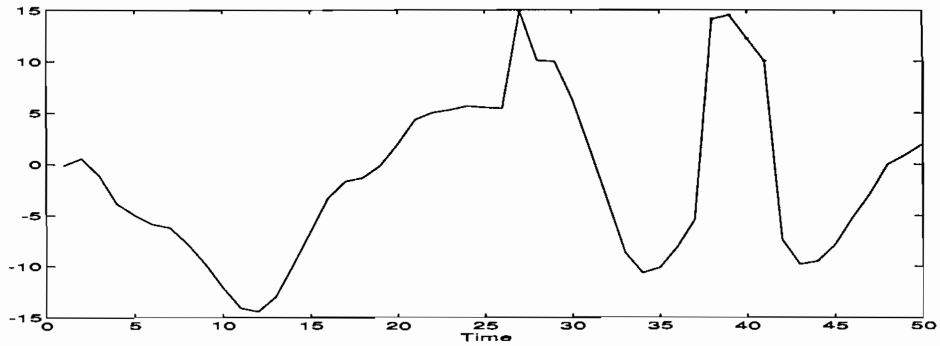



**Figure 1:** AR(3) artificial time series. Five outliers of sizes 7, 20, 20, 17 and 15 have been introduced in the positions $t = 27, 38, 39, 40$ and $41$.

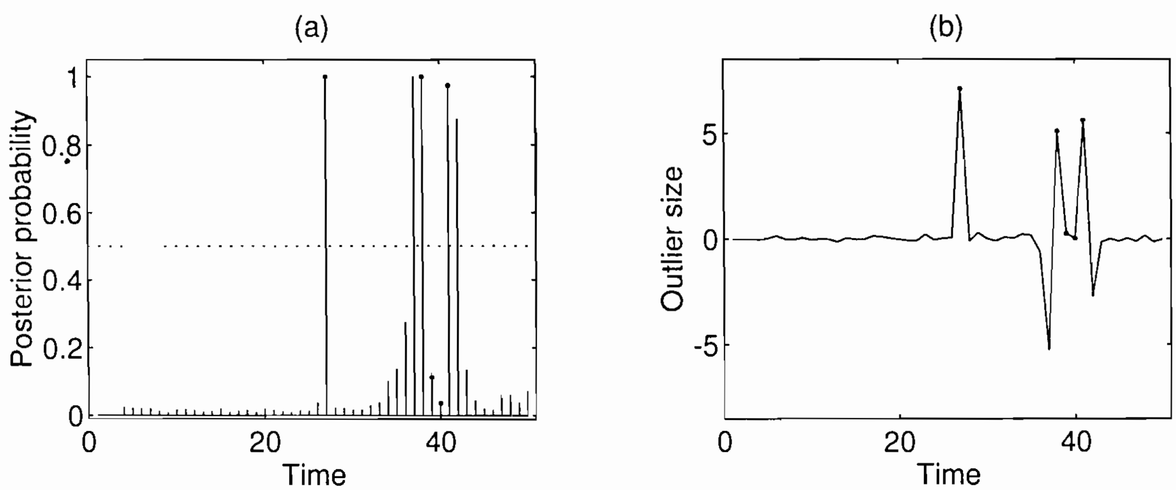



**Figure 2:** Results of the Gibbs sampler after 31,984 iterations for the artificial time series with five outliers: (a) posterior probabilities for each data point to be outlier; (b) posterior mean estimates of the outlier sizes for each data.

The paper is organized as follows. Section 2 reviews the application of the standard Gibbs sampler to outlier identification in an autoregressive time series. Section 3 proposes a new adaptive Gibbs algorithm to detect outlier patches. The conditional posterior distributions of blocks of observations are obtained and used to expedite the convergence of the algorithm. Section 4 illustrates the performance of the proposed procedure using several examples.

## 2. OUTLIER DETECTION IN AN AUTOREGRESSIVE TIME SERIES

### 2.1. AR model with additive outliers

Let $\{x_t\}$ be an AR process of order $p$, say

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t,$$

where $\{a_t\}$ is defined as in the Introduction and the polynomial $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ has no zeros inside the unit circle. In practice, we observe $\boldsymbol{y} = (y_1, \ldots, y_n)'$ such that

$$y_t = \delta_t \beta_t + x_t \qquad t = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$ is a binary random vector of outlier indicators; that is, $\delta_t = 1$ if the $t$-th observation is an outlier, and $\delta_t = 0$ otherwise. When $y_t$ is an outlier, $\beta_t$ denotes the size or magnitude of the outlier.

For simplicity, assume that $x_1, \ldots, x_p$ are fixed and $x_t = y_t$ for $t = 1, \ldots, p$ (i.e. no outliers in the first $p$ observations). Define $\boldsymbol{X}_t = (1, x_{t-1}, \ldots, x_{t-p})'$ and $\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_p)'$. The observed series can be expressed as a multiple regression model given by

$$y_t = \delta_t \beta_t + \boldsymbol{\phi}' \boldsymbol{X}_t + a_t \qquad t = p+1, \ldots, n. \tag{2.2}$$

It is understood here that the outlier indicator vector becomes $\boldsymbol{\delta} = (\delta_{p+1}, \ldots, \delta_n)'$ and the outlier size vector is $\boldsymbol{\beta} = (\beta_{p+1}, \ldots, \beta_n)'$. We also assume that the prior probability of being an outlier is the same for all observations, namely $P(\delta_t = 1) = \alpha$ for $t = p+1, \ldots, n$.

Abraham and Box (1979) obtained the posterior distributions for the parameters in the model (2.2), using the Jeffrey's reference prior distribution. In this particular case, the conditional posterior distribution of $\phi$, given the outlier configuration $\boldsymbol{\delta}_r$, is a multivariate $t$-distribution, where $\boldsymbol{\delta}_r$ is one of the $2^{n-p}$ possible outlier configurations. Therefore, the posterior distribution of $\phi$ is a mixture of $2^{n-p}$ multivariate $t$ distributions given by

$$P(\phi \mid \boldsymbol{y}) = \sum w_r P(\phi \mid \boldsymbol{y}, \boldsymbol{\delta}_r), \tag{2.3}$$

where the summation is over all $2^{n-p}$ possible outlier configurations and the weight $w_r$ is the posterior distribution of $\boldsymbol{\delta}_r$, i.e. $w_r = P(\boldsymbol{\delta}_r \mid \boldsymbol{y})$. For such a model, we can identify the outliers using the posterior marginals of elements of $\boldsymbol{\delta}$,

$$p_t = P(\delta_t = 1 \mid \boldsymbol{y}) = \sum P(\boldsymbol{\delta}_{r_t} \mid \boldsymbol{y}), \qquad t = p+1, \ldots, n, \tag{2.4}$$

where the summation is now over the $2^{n-p-1}$ outlier configurations $\boldsymbol{\delta}_r$ with $\delta_t = 1$ and the posterior probabilities $P(\boldsymbol{\delta}_{r_t} \mid \boldsymbol{y})$ are given by

$$P(\boldsymbol{\delta}_{r_t} \mid \boldsymbol{y}) = \int P(\boldsymbol{\delta}_{r_t} \mid \boldsymbol{y}, \boldsymbol{\Psi}) P(\boldsymbol{\Psi} \mid \boldsymbol{y}) \, d\boldsymbol{\Psi}$$

where $\boldsymbol{\Psi} = (\phi, \sigma_a^2, \boldsymbol{\beta})$ and $P(\boldsymbol{\delta}_{r_t} \mid \boldsymbol{y}, \boldsymbol{\Psi}) \propto P(\boldsymbol{y} \mid \boldsymbol{\delta}_{r_t}, \boldsymbol{\Psi}) P(\boldsymbol{\delta}_{r_t})$. Moreover, the posterior distributions of the outlier magnitudes are mixtures of Student $t$ distributions

$$P(\beta_t \mid \boldsymbol{y}) = \sum w_r P(\beta_t \mid \boldsymbol{y}, \boldsymbol{\delta}_r), \qquad t = p+1, \ldots, n. \tag{2.5}$$

The computation of the posterior probabilities (2.3), (2.4) and (2.5) is intensive even when the sample size is small, because these probabilities are mixtures of $2^{n-p}$ or $2^{n-p-1}$ distributions. Thus, such an approach becomes infeasible when the sample size is moderate or large. Alternative approach must be sought.

## 2.2. Gibbs sampling for isolated outlier detection

McCulloch and Tsay (1994) proposed to compute the posterior distributions (2.3), (2.4) and (2.5) by Gibbs sampling. The procedure requires full conditional posterior distributions of each parameter in model (2.2) given all the other parameters. For ease in reference, we summarize the necessary conditional posterior distributions in the following proposition. These results were first obtained by McCulloch and Tsay (1994) using conjugate prior distributions for $\phi$ and $\sigma_a^2$. Here we use non-informative priors for the parameters. The prior distribution for the contamination parameter $\alpha$ is $Beta(\gamma_1, \gamma_2)$, with expectation $E(\alpha) = \gamma_1/(\gamma_1 + \gamma_2)$. The outlier indicator $\delta_t$ and the outlier magnitude $\beta_t$ are assumed *a priori* independent and time-invariant with $Bernoulli(\alpha)$ and $N(0, \tau^2)$ distributions, respectively.

PROPOSITION 1. *Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be a vector of observations generated by the model (2.2), where $x_t = y_t$ are fixed for $t = 1, \ldots, p$ and $x_t = y_t - \delta_t\beta_t$ for $t = p+1, \ldots, n$. Assume that the prior distributions are*

$$\delta_t \sim Bernoulli(\alpha), \qquad t = p+1, \ldots, n$$

*and*

$$P(\boldsymbol{\phi}, \sigma_a^2, \alpha, \boldsymbol{\beta}) \propto \sigma_a^{-2}\alpha^{\gamma_1 - 1}(1-\alpha)^{\gamma_2 - 1}\exp\left(-\frac{1}{2\tau^2}\sum_{t=1}^{n}\beta_t^2\right),$$

*where the hyperparameters $\gamma_1$, $\gamma_2$ and $\tau^2$ are known. Then, the full conditional posterior distributions are as follows:*

*1. The conditional posterior distribution of the AR parameter vector $\boldsymbol{\phi}$ is given by*

$$\boldsymbol{\phi} \mid \boldsymbol{y}, \sigma_a^2, \boldsymbol{\delta}, \boldsymbol{\beta} \sim N_{p+1}\left(\boldsymbol{\phi}^*, \sigma_a^2\, \boldsymbol{\Omega}_\phi\right),$$

*where*

$$\boldsymbol{\Omega}_\phi = \left(\sum_{t=p+1}^{n} \boldsymbol{X}_t \boldsymbol{X}_t'\right)^{-1}, \qquad \boldsymbol{\phi}^* = \boldsymbol{\Omega}_\phi \sum_{t=p+1}^{n} \boldsymbol{X}_t x_t. \tag{2.6}$$

*2. The conditional posterior distribution of the innovational variance $\sigma_a^2$ is*

$$\sigma_a^2 \mid \boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\beta} \sim Inverted-Gamma\left(\frac{n-p}{2}, \frac{1}{2}\sum_{t=p+1}^{n} a_t^2\right).$$

*3. The conditional posterior distribution of $\delta_j$ for $j = p+1, \ldots, n$ is Bernoulli with* *probability*

$$P(\delta_j = 1 \mid \boldsymbol{y}, \boldsymbol{\phi}, \sigma_a^2, \boldsymbol{\delta}_{(j)}, \boldsymbol{\beta}, \alpha) = \frac{\exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_j} e_t(1)^2\right)\alpha}{\exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_j} e_t(1)^2\right)\alpha + \exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_j} e_t(0)^2\right)(1-\alpha)}, \tag{2.7}$$

*where $\boldsymbol{\delta}_{(j)}$ is obtained from $\boldsymbol{\delta}$ by eliminating the component $\delta_j$, $T_j = \min(n, j+p)$ and $e_t(\delta_j) = x_t - \phi_0 - \sum_{i=1}^{p} \phi_i x_{t-i}$ is the residual at time $t$ when the series is corrected by the identified outliers. It is easy to see that $e_t(1) = e_t(0) + \pi_{t-j}\beta_j$, where $\pi_0 = -1$ and $\pi_j = \phi_j$ for $j = 1, \ldots, p$.*

*The posterior distributions of the outlier magnitudes $\beta_j$ for $j = p+1, \ldots, n$ are*

$$\beta_j \mid \boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\beta}_{(j)}, \sigma_a^2 \sim N(\delta_j \beta_j^*, \sigma_j^2), \tag{2.8}$$

*where $\boldsymbol{\beta}_{(j)}$ is obtained from $\boldsymbol{\beta}$ by eliminating the component $\beta_j$ and the parameters are*

$$\sigma_j^2 = \frac{\tau^2 \sigma_a^2}{\tau^2 \nu_{T_j-j}^2 \delta_j + \sigma_a^2}, \tag{2.9}$$

*with $\nu_{T_j-j}^2 = (1 + \phi_1^2 + \cdots + \phi_{T_j-j}^2)$, and*

$$\beta_j^* = \frac{\sigma_j^2}{\sigma_a^2}[e_j(0) - \phi_1 e_{j+1}(0) - \ldots - \phi_{T_j-j} e_{T_j}(0)]. \tag{2.10}$$

*4. The conditional posterior distribution of $\alpha$ only depends on the vector $\boldsymbol{\delta}$ and is*

$$\alpha \mid \boldsymbol{\delta} \sim Beta[\gamma_1 + (n-p)\bar{\delta}\ ,\ \gamma_2 + (n-p)(1-\bar{\delta})],$$

*where $(n-p)\bar{\delta} = \sum_{t=p+1}^{n} \delta_t$. Consequently, the conditional posterior mean of $\alpha$ can be expressed as a linear combination of the prior mean and the mean of the data, that is, $E(\alpha \mid \boldsymbol{\delta}) = \omega E(\alpha) + (1-\omega)\bar{\delta}$, where $\omega = (\gamma_1 + \gamma_2)/(\gamma_1 + \gamma_2 + n - p)$.*

The equation (2.7) has a simple interpretation. The hypothesis $\delta_j = 1$ (i.e. $y_j$ is an outlier) given the data only affects the residuals $e_j, \ldots, e_{T_j}$. Assuming that the parameters are known, we can judge the likelihood of this hypothesis by: (a) computing these residuals when $\delta_j = 1$, that is, calculating $e_t(1)$ for $t = j, \cdots, T_j$; (b) computing the residuals when $\delta_j = 0$, that is, calculating $e_t(0)$; and (c) comparing the two sets of residuals. Equation (2.7) is simply the usual way of comparing the likelihoods of the null and alternative hypotheses. The probability in (2.7) can be written as

$$P(\delta_j = 1 \mid \boldsymbol{y}, \boldsymbol{\phi}, \sigma_a^2, \boldsymbol{\delta}_{(j)}, \boldsymbol{\beta}, \alpha) = \left(1 + \frac{(1-\alpha)}{\alpha} F_{10}(j)\right)^{-1}, \tag{2.11}$$

where $F_{10}$ is the Bayes factor given by

$$F_{10}(j) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_j}; \delta_j = 0)}{f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_j}; \delta_j = 1)}, \tag{2.12}$$

$\boldsymbol{\theta}_{\delta_j}$ denotes all the model parameters but $\theta_j$, and the logarithm of the squared Bayes factor is

$$\log F_{10}^2(j) = \frac{1}{\sigma_a^2} \left( \sum_{t=j}^{T_j} e_t(1)^2 - \sum_{t=j}^{T_j} e_t(0)^2 \right). \tag{2.13}$$

Since the residuals are the one-step-ahead prediction errors, Equation (2.13) compares the sum of prediction errors in the periods $j, j+1, \ldots, T_j$ when the forecasts are evaluated under the null hypothesis of $\delta_j = 1$ with that under the alternative hypothesis of $\delta_j = 0$. This is equivalent to the Chow test (Chow, 1960) for structural changes when the variance is known.

When there is no prior information about the outlier magnitudes ($\tau^2 \to \infty$) and the observation $y_j$ is identified as an outlier, the conditional posterior mean of $\beta_j$ tends to $\hat{\beta}_j = \nu_{T_j-j}^{-2}[e_j(0) - \phi_1 e_{j+1}(0) - \ldots \phi_{T_j-j} e_{T_j}(0)]$ which is the least squares estimate when the parameters are known (Chang, Tiao and Chen, 1988). The conditional posterior variance given by equation (2.9) is the variance of the estimate $\hat{\beta}_j$. The conditional posterior mean in (2.10) can also be seen as a linear combination of the prior mean and the magnitude of the outlier estimated from the data (Peña, 1990). The magnitude estimate is the difference between the observation $y_j$ and the linear predictor of $y_j$ given the data that minimizes the mean squared error. Equation (2.10) can be expressed as

$$\beta_j^\star = \frac{\tau^2 \nu_{T_j-j}^2}{\tau^2 \nu_{T_j-j}^2 + \sigma_a^2} (y_j - \hat{y}_{j|n}) + \frac{\sigma_a^2}{\tau^2 \nu_{T_j-j}^2 + \sigma_a^2} \beta_0, \tag{2.14}$$

where $\beta_0$ is the prior mean of $\beta_j$, which is zero in this paper, and $\hat{y}_{j|n}$ is the conditional expectation of $y_j$ given all the data. For the AR($p$) model under study, the optimal linear predictor $\hat{y}_{j|n}$ is a combination of the $p$ past and future values of $y_j$, and it is given by

$$\hat{y}_{j|n} = \phi_0 \nu_{T_j-j}^{-2} \tilde{\pi}_{T_j-j} - \nu_{T_j-j}^{-2} \left( \sum_{i=1}^{p} \sum_{t=0}^{T_j-j-i} \pi_t \pi_{t+i} x_{j-i} + \sum_{i=1}^{T_j-j} \sum_{t=0}^{T_j-j-i} \pi_t \pi_{t+i} x_{j+i} \right), \qquad (2.15)$$

where $\tilde{\pi}_t = 1 - \phi_1 - \ldots - \phi_t$ for $t \leq p$. When the observations are not near the end of the data series (say $j + p < n$) the filter (2.15) is

$$\hat{y}_{j|n} = \phi_0 \nu_p^{-2} \tilde{\pi}_p - [1 - \rho^{\mathrm{D}}(B)] x_j,$$

where $\rho^{\mathrm{D}}(B) = \nu_p^{-2} \pi(B) \pi(B^{-1})$ is the autocorrelation generating function of the dual process

$$x_t^{\mathrm{D}} = \phi_0 \tilde{\pi}_p + a_t - \phi_1 a_{t-1} - \ldots - \phi_p a_{t-p} \qquad (2.16)$$

introduced by Cleveland (1972), where $\nu_p^2$ is the variance of the dual process. In general, using the truncated autoregressive polynomial $\pi_{T_j-j}(B) = (1 - \pi_1 B - \ldots - \pi_{T_j-j} B^{T_j-j})$ and the "truncated" variance of the dual process, $\nu_{T_j-j}^2 = (1 + \pi_1^2 + \ldots + \pi_{T_j-j}^2)$, the estimate (2.15) can be written as a function of the "truncated" autocorrelation generating function $\rho_{T_j-j}^{\mathrm{D}}(B) = \nu_{T_j-j}^{-2} \pi_p(B) \pi_{T_j-j}(B^{-1})$ of the dual process. Therefore, for any period $j$ the optimal linear predictor of $y_j$ is

$$\hat{y}_{j|n} = \phi_0 \nu_{T_j-j}^{-2} \tilde{\pi}_{T_j-j} - [1 - \rho_{T_j-j}^{\mathrm{D}}(B)] x_j.$$

Let $\boldsymbol{\theta} = (\boldsymbol{\phi}, \sigma_a^2, \boldsymbol{\delta}, \boldsymbol{\beta}, \alpha)'$ be the vector of unknown parameters in model (2.2). One can use the results of Proposition 1 to draw Gibbs sampler for $\boldsymbol{\theta}$. These draws are easy as they are from the well-known Multivariate Normal, Inverted Gamma or Beta distributions. However, as demonstrated by the simple example in Introduction, such a Gibbs sampling procedure may fare poorly when the additive outliers appear in patches.

## 3. OUTLIER PATCH DETECTION

In this section, we propose a procedure to detect patches of additive outliers in an AR process. Our procedure is motivated by failure of the usual Gibbs sampler in handling such outliers and is aimed at improving the usual Gibbs sampler. The proposed procedure consists of two Gibbs runs. In the first run, the usual Gibbs sampling of Section 2 is applied to the data. The results of this Gibbs run are then used to implement the second Gibbs sampling that is adaptive in treating identified outliers and in using block interpolation to reduce possible masking and swamping effects.

As shown by the example in Introduction, when we have a patch of additive outliers with similar sizes, the beginning and the end observations of the patch are less affected by the masking effects whereas observations in the middle of the patch are subjected to substantial masking effects. Furthermore, "good" data points next to the outlying patch may be subjected to swamping effects. Such information is valuable in identifying the beginning and the end of a possible outlier patch, and suggests an adaptive way to determine the sizes of block interpolation for identifying consecutive outliers.

Another useful information resulting from the first Gibbs sampler is the estimated outlier sizes. These sizes can be used as the prior means of the outlier magnitudes $\beta_j$ in the second Gibbs sampler, resulting in an adaptive method to better estimate the sizes of identified outliers. This is in a sharp contrast with the first Gibbs sampler for which no prior information is available about outlier sizes so that the prior mean of $\beta_j$ is zero for all $j$. By Equation (2.14), the conditional posterior mean $\beta_j^*$ is a linear combination of the prior mean and the least squares estimate of $\beta_j$. A more informative prior mean can go a long way to improve the estimation of $\beta_j$ via Gibbs sampling.

A key feature of the proposed second Gibbs sampling is block interpolation. It enables us to draw outlier parameters jointly for possible outlying patches. An advantage of such groupings is that they speedup the convergence of the Gibbs sampler to the joint posterior distribution, especially when the parameters involved are highly correlated; see Liu, Wong and Kong (1994). When Gibbs draws are from a joint distribution of highly correlated parameters, the movements from a Gibbs iteration to the next one are in the principal components direction of the parameter space instead of parallel to the coordinate axes.

In what follows, we divide the detail of the proposed procedure into subsections. Our discussion focuses on the second Gibbs run and assumes that results of the first Gibbs run are available. For ease in reference, let $\hat{\boldsymbol{\phi}}^{(S)}$, $\hat{\sigma}_a^{(S)}$, $\hat{\boldsymbol{p}}^{(S)}$, and $\hat{\boldsymbol{\beta}}^{(S)}$ be the posterior means based on the last $R$ iterations of the first Gibbs run which uses $S$ iterations, where the $j$-th element of $\hat{\boldsymbol{p}}^{(S)}$ is $\hat{p}_{p+j}^{(S)}$ which is the posterior probability that $y_{p+j}$ is an outlier.

### 3.1. Location and joint estimation of outlier patches

The biases in $\hat{\boldsymbol{\beta}}^{(S)}$ induced by the masking effects of multiple outliers come from several sources. Two main sources are (a) drawing values of $\beta_j$ one by one and (b) the misspecification of the prior mean of $\beta_j$, which is fixed to zero. Due to the dynamic relation between observations in a time series, one-by-one drawing overlooks the dependence between parameters. For an AR($p$) process, an additive outlier affects $p+1$ residuals and the usual interpolation (or filtering) involves $p$ observations before and after the time index of interest. Consequently, an additive outlier affects the conditional posterior distributions of $2p+1$ observations; see equations (2.7) and (2.10). Chen and Liu (1993) pointed out that estimates of outlier magnitudes computed separately can differ markedly from those obtained from joint estimation. The situation becomes more serious in the presence of $k$ consecutive additive outliers for which the outliers affect $2p+k$ observations. To overcome this difficulty, we make use of the results of the first Gibbs sampler to identify possible locations and block sizes of outlier patches.

The tentative specification of locations and block sizes of outlier patches is done by a forward-backward search using a window around the outliers identified by the first Gibbs run. Let $c_1$ be a critical value between 0 and 1 so that any observation with posterior probability of being an outlier exceeds it would be classified as an "identified" outlier. That is, $y_j$ is identified as an outlier if $\hat{p}_j^{(S)} > c_1$. Typically, we use $c_1 = 0.5$. Let

$\{t_1, \cdots, t_m\}$ be the collection of time indexes of outliers identified by the first Gibbs run.

Turn to specification of patch size. Because of possible masking and swamping effects, we select another critical value $c_2$ with $c_2 \leq c_1$ to specify the beginning and end points of possible outlier patch associated with an identified outlier. In addition, to use the common sense that the number of outliers in a sample cannot be too large relatively to the sample size, we select a window of length $2hp$ to search for the boundary points of a possible outlier patch. Consider specifically the identified outlier $y_{t_i}$. First, we check the $hp$ observations *before* $y_{t_i}$ and compare their posterior probabilities $\hat{p}_j^{(S)}$ with $c_2$. Any point within the window with $\hat{p}_j^{(S)} > c_2$ is regarded as a possible beginning point of an outlier patch associated with $y_{t_i}$. We then select the farthest point from $y_{t_i}$ as the beginning point of the outlier patch. Denote the point by $y_{t_i-k_i}$. Second, do the same for the $hp$ observations *after* $y_{t_i}$ and select the farthest point from $y_{t_i}$ with $\hat{p}_j^{(S)} > c_2$ as the end point of the outlier patch. Denote the end point by $y_{t_i+v_i}$. Finally, combine the two blocks to form a tentative candidate for outlier patch associated with $y_{t_i}$. Denote the patch by $(y_{t_i-k_i}, \cdots, y_{t_i+v_i})$ the length of which is $v_i - k_i + 1$.

Finally, consider jointly all the identified outlier patches for further refinement. First, any overlapping or consecutive patches should be merged to form a larger patch. Secondly, if the total number of outliers is greater than $n/2$, where $n$ is the sample size, then one should increase $c_2$ and re-specify possible outlier patches. Thirdly, if increasing $c_2$ cannot sufficiently reduce the total number of outliers, one should decrease the window size by choosing a smaller $h$ and re-specify outlier patches.

In summary, the proposed approach to tentatively specify outlier patchers is as follows:

1. Choose $c_1$ and identify isolated outlying observations $y_{t_i}$ using $\hat{p}_{t_i}^{(S)} > c_1$. Denote the time indexes of identified outliers by $\{t_1, \ldots, t_m\}$.
2. Specify $h$ and $c_2$, where $c_2 \leq c_1$. For each $y_{t_i}$, identify a possible outlier patch $(y_{t_i-k_i}, \cdots, y_{t_i+v_i})$ using the procedure mentioned above with $c_2$ and $hp$ window.

3. Merge any overlapping or consecutive tentative patches.
4. If the total number of outliers is greater than $n/2$, then increase $c_2$ and go to Step 2. If increasing $c_2$ cannot sufficiently reduce the total number of outliers, increase $h$ and go to Step 2. If the total number of outliers is less than $n/2$, stop the search for outlier patches and perform the second Gibbs sampling.

### 3.2. Conditional distributions for blocks of observations

With outlier patches tentatively specified, it is best to draw Gibbs samples jointly within each patch. To this end, we consider in this subsection the conditional posterior distribution of a block of $k$ observations in a time series. Suppose that a patch of $k$ outliers starting at time index $j$ is identified. Let $\boldsymbol{\delta}_{j,k} = (\delta_j, \ldots, \delta_{j+k-1})'$ and $\boldsymbol{\beta}_{j,k} = (\beta_j, \ldots, \beta_{j+k-1})'$ be the vectors of outlier indicators and magnitudes, respectively, for the patch. Partitioning the parameter vector of the model as

$$\boldsymbol{\theta}_{\mathrm{B}} = (\boldsymbol{\phi}, \sigma_a^2, \delta_{p+1}, \ldots, \boldsymbol{\delta}'_{j,k}, \delta_{j+k}, \ldots, \delta_n, \beta_{p+1}, \ldots, \boldsymbol{\beta}'_{j,k}, \beta_{j+k}, \ldots, \beta_n, \alpha)',$$

we need the conditional posterior distributions of the parameters given the others. Most of these conditional posterior distributions are given in Proposition 1, but it remains to derive those for $\boldsymbol{\delta}_{j,k}$ and $\boldsymbol{\beta}_{j,k}$. We give the results in Theorems 1 and 2. The proofs are given in the Appendix.

THEOREM 1. *Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be a vector of observations from model (2.2). Assume independent prior distributions Bernoulli($\alpha$) for $\delta_j, \ldots, \delta_{j+k-1}$ and $j > p$. Then, the conditional posterior distribution of $\boldsymbol{\delta}_{j,k}$ given the sample and other parameters $\boldsymbol{\theta}_{\delta_{j,k}} = (\boldsymbol{\phi}, \boldsymbol{\delta}_{p+1,j-1}, \boldsymbol{\delta}_{j+k,n}, \boldsymbol{\beta}, \sigma_a^2, \alpha)'$ is*

$$P(\boldsymbol{\delta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\delta_{j,k}}) = C\ \alpha^{\boldsymbol{s}_{j,k}}(1-\alpha)^{k-\boldsymbol{s}_{j,k}}\ \exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_{j,k}} e_t(\boldsymbol{\delta}_{j,k})^2\right) \tag{3.17}$$

*where $\boldsymbol{s}_{j,k} = \sum_{t=j}^{j+k-1} \delta_t$, $T_{j,k} = \min\{n, j+k+p-1\}$, $C$ is a normalization constant so that the total probability of the $2^k$ possible configurations of $\boldsymbol{\delta}_{j,k}$ is one, and $e_t(\boldsymbol{\delta}_{j,k}) = x_t - \phi_0 - \sum_{i=1}^{p} \phi_i x_{t-i}$ is the residual at time $t$ when the series is corrected by the identified outliers not included in the interval $[j, j+k-1]$.*

THEOREM 2. *Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be a vector of observations from model (2.2). Assume independent prior distributions $N(\beta_0, \tau^2)$ for the elements of the vector $\boldsymbol{\beta}_{j,k}$. Then, the conditional posterior distribution of $\boldsymbol{\beta}_{j,k}$ given the sample and other parameters is*

$$\boldsymbol{\beta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\delta}, \sigma_a^2, \alpha \sim N_k\left(\boldsymbol{\beta}_{j,k}^{\star}, \boldsymbol{\Omega}_{j,k}\right), \tag{3.18}$$

*where*

$$\boldsymbol{\Omega}_{j,k} = \left(\boldsymbol{D}_{j,k}\left(\frac{1}{\sigma_a^2}\sum_{t=j}^{T_{j,k}} \boldsymbol{\Pi}_{t-j}\boldsymbol{\Pi}_{t-j}'\right)\boldsymbol{D}_{j,k} + \frac{1}{\tau^2}\boldsymbol{I}\right)^{-1} \tag{3.19}$$

*and*

$$\boldsymbol{\beta}_{j,k}^{\star} = \boldsymbol{\Omega}_{j,k}\left(-\frac{1}{\sigma_a^2}\sum_{t=j}^{T_{j,k}} e_t(\boldsymbol{0})\boldsymbol{D}_{j,k}\boldsymbol{\Pi}_{t-j} + \frac{1}{\tau^2}\beta_0\right), \tag{3.20}$$

*where $T_{j,k} = \min\{n, j+k+p-1\}$ and the residual $e_t(\boldsymbol{0})$ is defined in Theorem 1, $\boldsymbol{D}_{j,k}$ is a $k \times k$ diagonal matrix with elements $\delta_j, \ldots, \delta_{j+k-1}$, and $\boldsymbol{\Pi}_t = (\pi_t, \pi_{t-1}, \ldots, \pi_{t-k+1})'$ is a $k \times 1$ vector, where $\pi_0 = -1$, $\pi_i = \phi_i$ for $i = 1, \ldots, p$, and $\pi_i = 0$ for $i < 0$ or $i > p$.*

Define the matrices $\boldsymbol{W}_1 = \sigma_a^{-2}\boldsymbol{\Omega}_{j,k}\left(\boldsymbol{D}_{j,k}\sum_{t=j}^{T_{j,k}} \boldsymbol{\Pi}_{t-j}\boldsymbol{\Pi}_{t-j}'\boldsymbol{D}_{j,k}\right)$ and $\boldsymbol{W}_2 = \tau^{-2}\boldsymbol{\Omega}_{j,k}$. Then $\boldsymbol{\beta}_{j,k}^{\star}$ can be written as

$$\boldsymbol{\beta}_{j,k}^{\star} = \boldsymbol{W}_1 \cdot \tilde{\boldsymbol{\beta}}_{j,k} + \boldsymbol{W}_2 \cdot \boldsymbol{\beta}_0,$$

where $\boldsymbol{W}_1 + \boldsymbol{W}_2 = \boldsymbol{I}$, implying that the mean of the conditional posterior distribution of $\boldsymbol{\beta}_{j,k}$ is a linear combination of the prior mean vector of $\boldsymbol{\beta}_0$ and the least squares estimate (or the maximum likelihood estimate) of the outlier magnitudes for a patch of data, that is

$$\tilde{\boldsymbol{\beta}}_{j,k} = \left(\boldsymbol{D}_{j,k}\sum_{t=j}^{T_{j,k}} \boldsymbol{\Pi}_{t-j}\boldsymbol{\Pi}_{t-j}'\boldsymbol{D}_{j,k}\right)^{-1}\left(-\sum_{t=j}^{T_{j,k}} e_t(\boldsymbol{0})\boldsymbol{D}_{j,k}\boldsymbol{\Pi}_{j,k}\right). \tag{3.21}$$

Maravall and Peña (1997) proved that, when $\delta_t = 1$ the estimate in (3.21) is equivalent to the vector of differences between the observations $(y_j, \ldots, y_{j+k-1})$ and the predictions $\hat{y}_t = E(y_t \mid y_1, \ldots, y_{j-1}, y_{j+k}, \ldots, y_n)$ for $t = j, \ldots, j+k-1$. The matrix $\boldsymbol{\Pi} =$

$\sum_{t=j}^{T_{j,k}} \mathbf{\Pi}_{t-j}\mathbf{\Pi}'_{t-j}$ is the $k \times k$ submatrix of the "truncated" autocovariance generating matrix of the dual process in (2.16). Specifically,

$$\mathbf{\Pi} = \begin{pmatrix} \nu^2_{T_{j,k}-j} & \gamma^{\mathrm{D}}_{1,T_{j,k}-j-1} & \cdots & \gamma^{\mathrm{D}}_{k-1,T_{j,k}-j-k+1} \\ \gamma^{\mathrm{D}}_{-1,T_{j,k}-j} & \nu^2_{T_{j,k}-j-1} & \cdots & \gamma^{\mathrm{D}}_{k-2,T_{j,k}-j-k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma^{\mathrm{D}}_{-k+1,T_{j,k}-j} & \gamma^{\mathrm{D}}_{-k+2,T_{j,k}-1} & \cdots & \nu^2_{T_{j,k}-j-k+1} \end{pmatrix},$$

where $\gamma^{\mathrm{D}}_{i,j} = \nu^2_j \rho^{\mathrm{D}}_{i,j}$, $\nu^2_j$ is the "truncated" variance of the dual process and $\rho^{\mathrm{D}}_{i,j}$ is the coefficient of $B^i$ in the "truncated" autocorrelation generating function of the dual process, i.e. $\rho^{\mathrm{D}}_j(B) = \nu_j^{-2}\pi_p(B)\pi_j(B^{-1})$.

### 3.3. The Second Gibbs Sampling

Using the results of the previous subsections, we now consider the second and adaptive Gibbs run of the proposed procedure. First of all, the results of the first Gibbs run provide useful information to start the second Gibbs sampling and to specify prior distributions of the parameters. For example, the starting values of $\delta_t$ for the second Gibbs sampling are given as follows: $\delta_t^{(0)} = 1$ if $\hat{p}_t^{(\mathrm{S})} > 0.5$ or if $y_t$ belongs to an identified outlier patch; otherwise, $\delta_t^{(0)} = 0$. To reduce the bias in estimating $\beta_t^\star$, the prior distributions of $\beta_t$ is given below:

a) If $y_t$ is identified as an isolated outlier the prior distribution of $\beta_t$ is $N(\hat{\beta}_t^{(\mathrm{S})}, \tau^2)$, where $\hat{\beta}_t^{(\mathrm{S})}$ is the Gibbs estimate of $\beta_t$ from the first Gibbs run.

b) If $y_t$ belongs to any outlier patch the prior distribution of $\beta_t$ is $N(\tilde{\beta}_t^{(\mathrm{S})}, \tau^2)$, where $\tilde{\beta}_t^{(\mathrm{S})}$ is the conditional posterior mean given in (3.21).

c) If $y_t$ does not belong to any outlier patch, nor an isolated outlier, then the prior distribution of $\beta_t$ is $N(0, \tau^2)$.

For each outlier patch, the results of Theorems 1 and 2 are used to draw $\boldsymbol{\delta}_{j,k}$ and $\boldsymbol{\beta}_{j,k}$ in the second Gibbs sampling, which is also run for $S$ iterations, but only the results of the last $R$ iterations are used to make inference. In practice, we use

$R = 1,000$ and the number $S$ can be determined by any sequential method proposed in the literature to monitor the convergence of Gibbs sampling. In this paper, we use a method that can be easily implemented. It is based on monitoring the estimates of outlying probability for each data point. Specifically, for $i$ iterations, where $i > 2,000$, we compute $\hat{p}_{p+1}^{(i)}, \ldots, \hat{p}_n^{(i)}$ using the draws from the last $i - 1,000$ iterations. The stopping-rule is the minimum number of iterations $S = i + R$ such that $\mid \hat{p}_t^{(i)} - \hat{p}_t^{(i-1)} \mid < \epsilon$ holds for all $t = p+1, \ldots, n$, where $\epsilon$ is a prespecified small positive number. In other words, we monitor the behavior of outlying probabilities to insure convergence of the adaptive Gibbs sampling.

An alternative procedure for handling outlier patches is to use the ideas of Bruce and Martin (1989). Select a positive integer $k$ in the interval $[1, n/2]$ as the maximum length of outlier patches in the data. Start the Gibbs sampler with $n - k - p$ parallel trials. In the $j$th trial, for $j = 1, \ldots, n - k - p$, the points at $t = p + j$ to $p + k + j$ are assigned initially as outliers. For other data points, use the usual initial outlier assignment. In application, one can use several different $k$ values. However, such a procedure requires intensive computation. On the other hand, the procedure proposed in this paper takes advantage of the first Gibbs sampler which can substantially reduce the computational burden.

## 4. APPLICATIONS

To illustrate the new adaptive Gibbs sampling algorithm we re-analyze the simulated time series of Section 1 and consider a real example. We compare the results of the usual Gibbs sampling, referred to as the standard Gibbs sampling, with those of the adaptive Gibbs sampling to see the efficacy of the latter algorithm. The real example demonstrates the applicability and effectiveness of the adaptive Gibbs sampling. It also shows that patches of outliers occur often in applications.

### 4.1. Simulated Data Revisited

As shown in Figure 2, the standard Gibbs sampling can easily detect the isolated outlier at $t = 27$ of the simulated AR(3) example, but it fails to identify the outlier patch in the period $t = 38$ to 42. To compare the results between the standard and adaptive Gibbs samlpings, we choose the hyperparameters $\gamma_1 = 5$, $\gamma_2 = 95$ and $\tau = 3$, implying that the contamination parameter has a prior mean $\alpha_0 = 0.05$, and the prior standard deviation of $\beta_t$ is three times the residual standard deviation. Using $\epsilon = 10^{-5}$ to monitor convergence, we obtained $S = 31,984$ iterations for the first Gibbs sampling and $S = 23,674$ iterations for the second and adaptive Gibbs sampling. All of the parameter estimates reported are the sample means of the last $R = 1,000$ iterations. For specifying the location of an outlier patch, we choose the criterion parameters $c_1 = 0.5$ and $c_2 = 0.3$, and the window length $2p$ to search for the boundary points of the possible outlier patches, where $p = 3$ is the autoregressive order of the series. Additional checking confirms that the results are stable with minor modifications of these parameter values.

The results of the first run (standard Gibbs sampling) are shown in Figure 2 and summarized in Tables 2 and 3. As before, the procedure indicates a possible patch of outliers from $t = 37$ to 42. In the second run (adaptive Gibbs sampling) the initial conditions and the prior distributions are specified by the proposed adaptive procedure. The posterior probability of outlier for each data point, $\hat{p}_t^{(S)}$, is shown in Figure 3(a). Clearly, the adaptive Gibbs sampling successfully specifies all the outliers, and there are no swamping or masking effects. In Table 3, we compare the estimates of the sizes of outliers between the adaptive and the standard Gibbs sampling, and in Table 2 we make the same comparison for the estimated parameter values. Both tables demonstrate clearly the efficacy and added values of the adaptive Gibbs sampling.

| Parameter | $\phi_0$ | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\sigma_a^2$ |
|---|---|---|---|---|---|
| True Value | 0 | 2.1 | -1.46 | 0.33 | 1 |
| Standard Gibbs | -0.09 | 2.16 | -1.82 | 0.58 | 2.15 |
| Adaptive Gibbs | -0.12 | 2.13 | -1.65 | 0.45 | 1.16 |

**Table 2:** True parameter values and estimates obtained by the standard and the adaptive Gibbs sampling algorithms.

| Time index | 27 | 37 | 38 | 39 | 40 | 41 | 42 |
|---|---|---|---|---|---|---|---|
| True Size | 7 | 0 | 20 | 20 | 17 | 15 | 0 |
| Standard Gibbs | 7.09 | -5.24 | 5.11 | 0.02 | 0.01 | 5.61 | -2.71 |
| Adaptive Gibbs | 7.28 | -0.09 | 17.35 | 16.78 | 15.01 | 14.73 | 0.02 |

**Table 3:** True outlier sizes and their estimates by the standard and the adaptive Gibbs sampling algorithms.

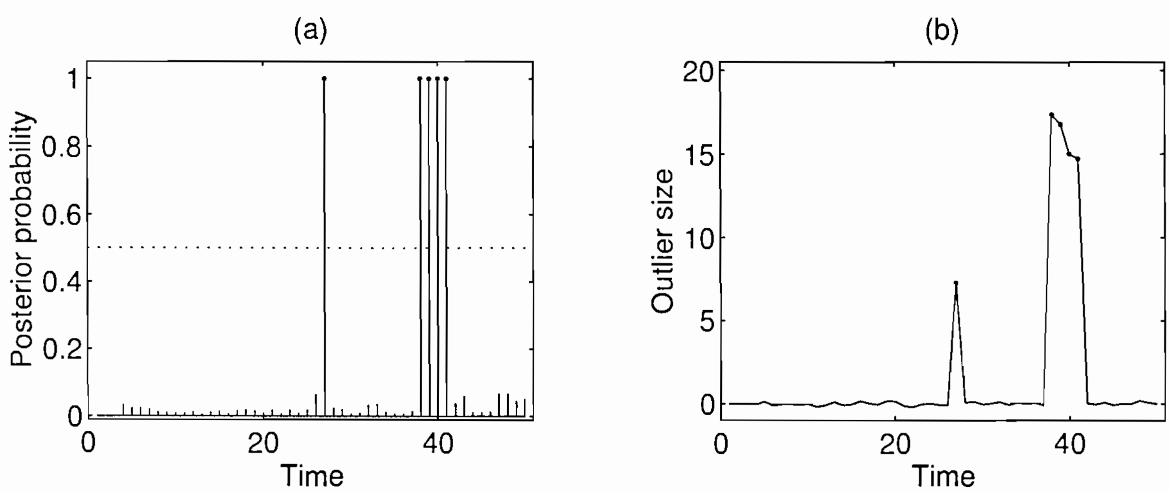



**Figure 3:** Adaptive Gibbs sampling results with 23,674 iterations for the artificial time series with five outliers. Only the last 1,000 iterations are used to make inference: (a) posterior probabilities for each data point to be outlier; (b) posterior mean estimates of the outlier sizes for each data.

| Parameter | $\phi_0$ | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\sigma_a$ |
|---|---|---|---|---|---|
| Initial model | 5.98 | 0.62 | 0.19 | 0.15 | 0.091 |
| Standard Gibbs | 0.19 | 0.82 | 0.19 | -0.05 | 0.062 |
| Adaptive Gibbs | 0.16 | 0.87 | 0.20 | -0.10 | 0.054 |

Table 4: Estimated parameter values with the initial model and those obtained by the standard and the adaptive Gibbs sampling algorithms.

## 4.2. A Real Example

We consider the data of monthly U.S. Industry-unfilled orders for radio and TV, in millions of dollars, which has been previously studied by Bruce and Martin (1989), among others. We use the logged series from January 1958 to October 1980, and shall focus on the seasonally adjusted series, where the seasonal component was removed by the well-known X11-ARIMA procedure. The seasonally adjusted series is shown in Figure 4. An AR(3) model is fitted to the data and the estimated parameter values are given in the first row of Table 4. The residual plot of this model shown in Figure 5 indicates some possible isolated outliers and outlier patches, especially in the latter part of the series.

The hyperparameters needed to run the adaptive Gibbs algorithm are set by the same criteria as those of the simulated example: $\gamma_1 = 5$, $\gamma_2 = 95$, $\tau = 3\sigma_a = 0.273$. In this particular instance, the stopping criterion $\epsilon = 10^{-5}$ is achieved by 53,720 iterations in the first run and by 72,434 iterations in the second Gibbs sampling. As before, to specify possible outlier patches prior to running the adaptive Gibbs sampling, the window width is set to twice of the AR order, $c_1 = 0.5$ and $c_2 = 0.3$. In addition, we have assumed that the maximum length of an outlier patch is 11 months, just below one year.

Using 0.5 as the cut-off posterior probability to identify outliers, we summarize the results of the standard Gibbs sampling and the adaptive Gibbs sampling in Ta-


**Figure 4:** Seasonally adjusted series of the logarithm of U.S. Industry-unfilled orders for radio and TV.

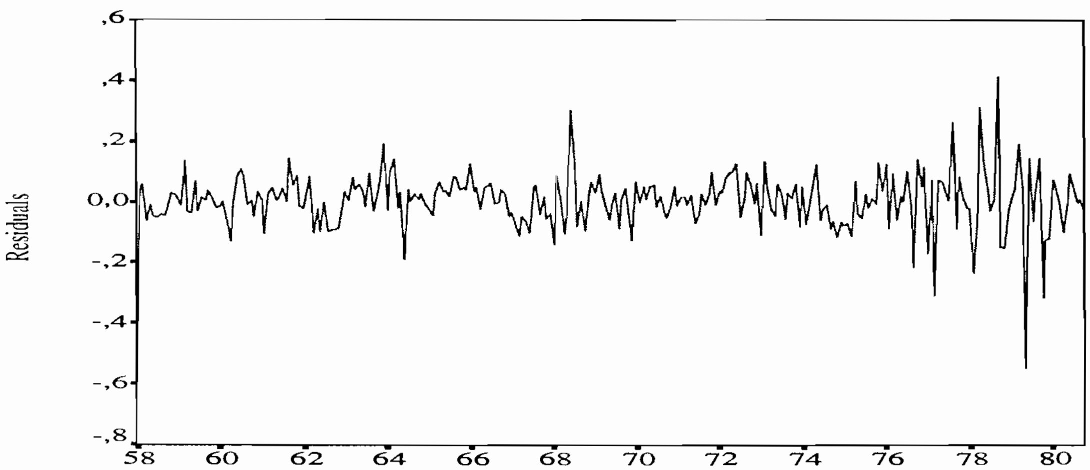



**Figure 5:** Residual plot for an AR(3) model fitted to the seasonal adjusted series of logarithms of U.S. Industry-unfilled orders for radio and TV

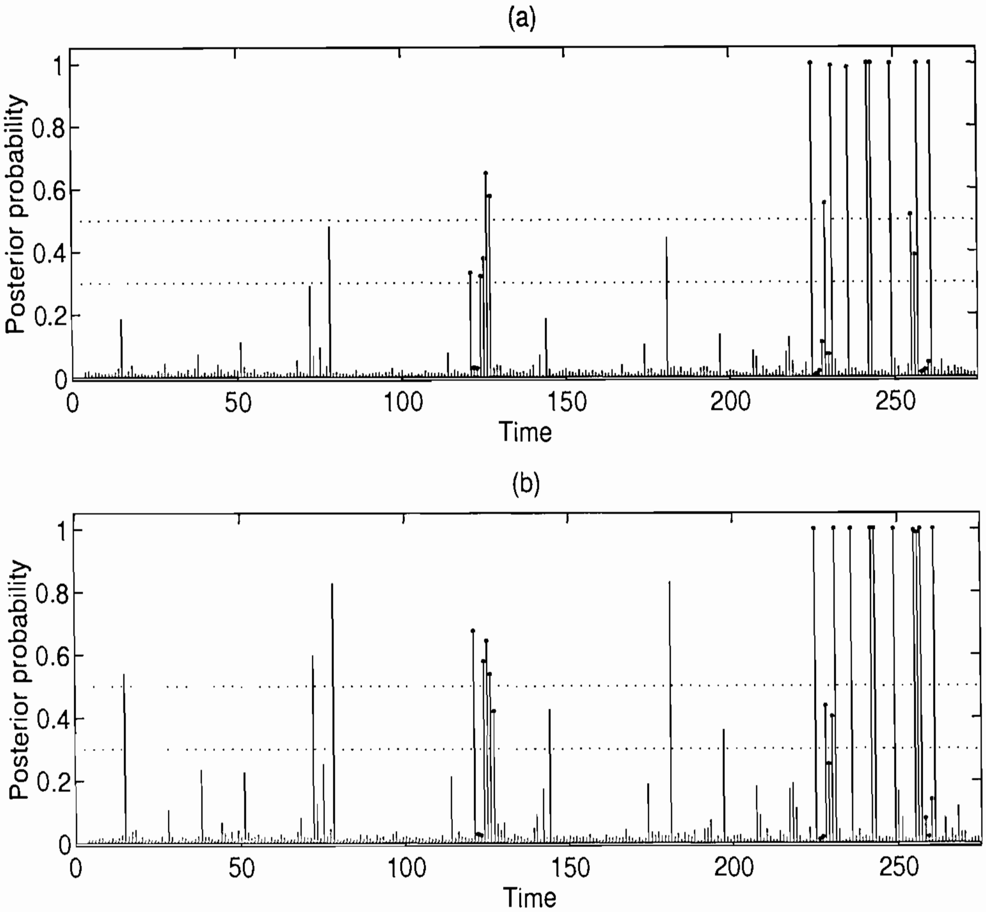



**Figure 6:** Posterior probability for each data point to be outlier with: (a) standard Gibbs sampling (first run); and (b) adaptive Gibbs sampling (second run).

bles 5 and 6). The standard Gibbs algorithm identifies 12 data points as outliers. They correspond to 8 isolated outliers and two outlier patches both of length 2. The two outlier patches are 6-7/1968 and 2-3/1978. The posterior probability of outlier for each data point is shown in Figure 6(a), and the estimated outlier sizes are given in Table 5. On the other hand, the second and adaptive Gibbs sampling specifies 18 data points as outliers. They consist of 10 isolated outliers, and 3 outlier patches of length 3, 2 and 3, respectively. The outlier patches are 4-6/1968, 2-3/1978, and 3-5/1979, respectively. The posterior probabilities of outlier based on results of this adaptive Gibbs sampling are presented in Figure 6(b), and the estimated outlier sizes and their outlying probabilities are summarized in Tables 5 and 6. Finally, Table 7 presents the detail of the posterior probabilities of outlier for each data point in the three detected outlier patches both for the standard and adaptive Gibbs sampling. It is interesting to contrast the two Gibbs sampling algorithms. First, for the possible outlier patch from January to July 1968, the two algorithms show dramatically different results. The standard Gibbs sampling only identifies the shorter patch 6-7/1968 as outliers. In contrast, the adaptive Gibbs sampling detects an isolated outlier at 1/68 and a longer outlier patch 4-6/1968, but removes 7/68 as an outlier. Therefore, we see both the masking and the swamping effects of multiple outliers in the standard algorithm. Secondly, within the possible outlier patch from September 1976 to March 1977, the adaptive algorithm slightly lowers the outlying probability for January 1977 so that it is no longer an outlier. Thirdly, consider the possible outlier patch from March to September 1979. The standard algorithm identifies two isolated outliers in April and September. On the other hand, the adaptive algorithm substantially increases the outlying posterior probabilities for March and April of 1979 and, hence, changes an isolated outlier into a patch of three outliers. The isolated outlier in September remains unchanged. Here the standard algorithm encounters severe masking effects.

The results of this example clearly demonstrate that (a) outlier patches occur frequently in practice and (b) similarly to the simulated example, the standard Gibbs sampling to outlier detection may encounter severe masking and swamping effects.

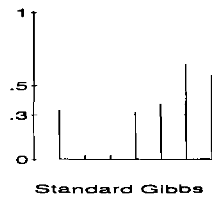
| Date | 3/59 | 12/63 | 6/74 | 1/68 | 4/68 | 5/68 | 6/68 | 7/68 | 1/73 | 9/76 |
|---|---|---|---|---|---|---|---|---|---|---|
| Standard Gibbs | 0.008 | 0.040 | -0.060 | -0.045 | -0.046 | -0.066 | 0.146 | 0.092 | -0.048 | -0.221 |
| Adaptive Gibbs | 0.066 | 0.075 | -0.116 | -0.139 | -0.165 | -0.173 | 0.149 | 0.143 | -0.108 | -0.211 |
| Date | 1/77 | 3/77 | 8/77 | 2/78 | 3/78 | 9/78 | 3/79 | 4/79 | 5/79 | 9/79 |
| Standard Gibbs | -0.082 | -0.237 | 0.203 | -0.266 | -0.271 | 0.366 | 0.092 | 0.072 | -0.410 | 0.265 |
| Adaptive Gibbs | -0.029 | -0.203 | 0.216 | -0.277 | -0.287 | 0.378 | 0.228 | 0.227 | -0.289 | 0.331 |

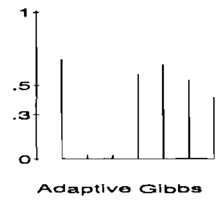
Table 5: Estimated outlier sizes by the standard and the adaptive Gibbs sampling algorithms.

| Date | 3/59 | 12/63 | 6/74 | 1/68 | 4/68 | 5/68 | 6/68 | 7/68 | 1/73 | 9/76 |
|---|---|---|---|---|---|---|---|---|---|---|
| Standard Gibbs | 0.186 | 0.290 | 0.478 | 0.331 | 0.321 | 0.377 | 0.648 | 0.576 | 0.442 | 0.999 |
| Adaptive Gibbs | 0.539 | 0.595 | 0.827 | 0.674 | 0.577 | 0.643 | 0.536 | 0.420 | 0.829 | 1.000 |
| Date | 1/77 | 3/77 | 8/77 | 2/78 | 3/78 | 9/78 | 3/79 | 4/79 | 5/79 | 9/79 |
| Standard Gibbs | 0.553 | 0.991 | 0.985 | 1.000 | 1.000 | 1.000 | 0.516 | 0.387 | 1.000 | 1.000 |
| Adaptive Gibbs | 0.252 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.995 | 0.987 | 1.000 | 1.000 |

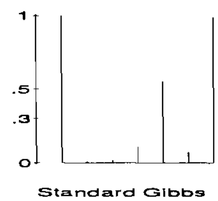
Table 6: Posterior probabilities for each data point to be outlier by the standard and the adaptive Gibbs sampling algorithms.

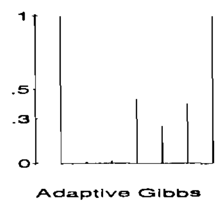
| Date | 1/68 | 2/68 | 3/68 | 4/68 | 5/68 | 6/68 | 7/68 |
|---|---|---|---|---|---|---|---|
| Standard Gibbs | 0.331 | 0.030 | 0.028 | 0.321 | 0.377 | 0.648 | 0.576 |
| Adaptive Gibbs | 0.674 | 0.029 | 0.025 | 0.577 | 0.643 | 0.536 | 0.420 |



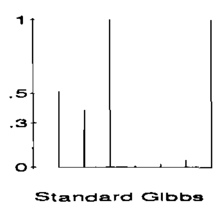




| Date | 9/76 | 10/76 | 11/76 | 12/76 | 1/77 | 2/77 | 3/77 |
|---|---|---|---|---|---|---|---|
| Standard Gibbs | 0.999 | 0.009 | 0.019 | 0.110 | 0.553 | 0.072 | 0.991 |
| Adaptive Gibbs | 1.000 | 0.012 | 0.019 | 0.436 | 0.252 | 0.402 | 1.000 |




| Date | 3/79 | 4/79 | 5/79 | 6/79 | 7/79 | 8/79 | 9/79 |
|---|---|---|---|---|---|---|---|
| Standard Gibbs | 0.516 | 0.387 | 1.000 | 0.015 | 0.022 | 0.046 | 1.000 |
| Adaptive Gibbs | 0.995 | 0.987 | 1.000 | 0.078 | 0.020 | 0.138 | 1.000 |



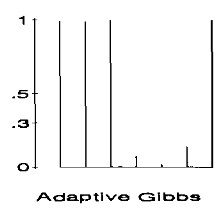




**Table 7:** Posterior probabilities for each data point to be outlier for the three larger possible outlier patches.

## ACKNOWLEDGEMENTS

A. Justel was financed by the European Commission in the Training and Mobility of Researchers Programme (TMR), when she was visiting C.O.R.E at Université Catholique de Louvain. D. Peña acknowledges support for this research provided by DGICYT (Spain), under grant No. PB96–0111. The work of R. Tsay is supported by the National Science Foundation and the Graduate School of Business, University of Chicago.

## REFERENCES


Abraham, B. and Box, G.E.P. (1979). "Bayesian analysis of some outlier problems in time series". *Biometrika,* 66, 229–236.

Bruce, A.G. and Martin, D. (1989). "Leave-k-out diagnostics for time series" (with discussion). *Journal of the Royal Statistical Society, B,* 51, 363-424.

Bustos, O.H. and Yohai, V.J. (1986). "Robust estimates for ARMA models". *Journal of the American Statistical Association,* 81, 155–168.

Chang, I. and Tiao, G.C. (1983). "Estimation of time series parameters in the presence of outliers". Technical Report 8, Statistics Research Center, University of Chicago.

Chang, I., Tiao, G.C. and Chen, C. (1988). "Estimation of time series parameters in the presence of outliers". *Technometrics,* 3, 193–204.

Chen, C. and Liu, L. (1993). "Joint estimation of model parameters and outlier effects in time series". *Journal of the American Statistical Association,* 88, 284–297.

Chow, G.C. (1960). "A test for equality between sets of observations in two linear regressions". *Econometrica,* 28, 591–605.

Cleveland, W.P. (1972). "The inverse autocorrelations of a time series and their applications". *Technometrics,* 14, 277- -298.

Denby, L. and Martin, D. (1979). "Robust estimation of the first-order autoregressive parameter". *Journal of the American Statistical Association,* 74, 140–146.

Fox, A.J. (1972). "Outliers in time series". *Journal of the Royal Statistical Society, B,* 34, 350–363.

Justel, A. and Peña, D. (1996). "Gibbs sampling will fail in outlier problems with strong masking". *Journal of Computational and Graphical Statistics,* 5, 176–189.

Liu, J.S., Wong, W.H. and Kong, A. (1994). "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes". *Biometrika*, 81, 27-40.

Maravall, A. and Peña, D. (1997). "Missing observations and additive outliers in time series models". *Advances in Statistical Analysis and Statistical Computing.* JAI Press (in press).

Martin, R.D., Samarov, A. and Vandaele, W. (1983). "Robust methods for ARIMA models". *Applied Time Series Analysis of Economic Data,* 153–169. Ed. A. Zellner, Bureau of the Census.

McCulloch, R.E. and Tsay R.S. (1994). "Bayesian analysis of autoregressive time series via the Gibbs sampler". *Journal of Time Series Analysis,* 15, 235–250.

Peña, D. (1987). "Measuring the importance of outliers in ARIMA models". *New Perspectives in Theoretical and Applied Statistics,* 109–118. Ed. Puri *et al.*, John Wiley.

Peña, D. (1990). "Influential observations in time series". *Journal of Business & Economic Statistics,* 8, 235–241.

Sánchez, M.J. and Peña, D. (1997). "The identification of multiple outliers in ARIMA models". Technical Report, Department of Statistics and Econometrics, Universidad Carlos III de Madrid.

Tsay, R.S. (1986). "Time series model specification in the presence of outliers", *Journal of the American Statistical Association,* 81, 132–141.

Tsay, R.S. (1988). "Outliers, level shifts, and variance change in time series". *Journal of Forecasting,* 7, 1–20.

## APPENDIX: PROOFS

*Proof of Theorem 1.* The conditional distribution of $\boldsymbol{\delta}_{j,k}$ given the sample and the other parameters is

$$P(\boldsymbol{\delta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\delta_{j,k}}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) \cdot \alpha^{\boldsymbol{S}_{j,k}}(1-\alpha)^{k-\boldsymbol{S}_{j,k}}. \tag{4.22}$$

The likelihood function can be factorized as

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) = f(\boldsymbol{y}_{p+1}^{j-1} \mid \boldsymbol{\theta}_{\delta_{j,k}}) \cdot f(\boldsymbol{y}_j^{T_{j,k}} \mid \boldsymbol{y}_{p+1}^{j-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) \cdot f(\boldsymbol{y}_{T_{j,k}+1}^{n} \mid \boldsymbol{y}_{p+1}^{T_{j,k}}, \boldsymbol{\theta}_{\delta_{j,k}}),$$

where $\boldsymbol{y}_j^k = (y_j, \ldots, y_k)'$. Only $f(\boldsymbol{y}_j^{T_{j,k}} \mid \boldsymbol{y}_{p+1}^{j-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k})$ depends on $\boldsymbol{\delta}_{j,k}$ and it is the product of the conditional densities:

$$\begin{aligned}
f(y_j \mid \boldsymbol{y}_{p+1}^{j-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \delta_j) &\propto \exp\left(-\frac{1}{2\sigma_a^2}(e_j(\mathbf{0}) - \delta_j\beta_j)^2\right) \\
&\vdots \\
f(y_{j+k-1} \mid \boldsymbol{y}_{p+1}^{j+k-2}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) &\propto \exp\left(-\frac{1}{2\sigma_a^2}(e_{j+k-1}(\mathbf{0}) - \delta_{j+k-1}\beta_{j+k-1} + \ldots + \pi_{k-1}\delta_j\beta_j)^2\right) \\
f(y_{j+k} \mid \boldsymbol{y}_{p+1}^{j+k-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) &\propto \exp\left(-\frac{1}{2\sigma_a^2}(e_{j+k}(\mathbf{0}) + \pi_1\delta_{j+k-1}\beta_{j+k-1} + \ldots + \pi_k\delta_j\beta_j)^2\right) \\
&\vdots \\
f(y_{T_{j,k}} \mid \boldsymbol{y}_{p+1}^{T_{j,k}-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) &\propto \exp\left(-\frac{1}{2\sigma_a^2}(e_{T_{j,k}}(\mathbf{0}) + \pi_{T_{j,k}-j-k+1}\delta_{j+k-1}\beta_{j+k-1} + \ldots + \right. \\
&\qquad \left. + \ \pi_{T_{j,k}-j}\delta_j\beta_j)^2\right).
\end{aligned}$$

Hence the likelihood function can be expressed as

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) \propto \exp\left(-\frac{1}{2\sigma_a^2}\left(\sum_{t=j}^{j+k-1}(e_t(\mathbf{0}) + \sum_{i=0}^{t-j}\pi_i\delta_{t-i}\beta_{t-i})^2 + \sum_{t=j+k}^{T_{j,k}}(e_t(\mathbf{0}) + \sum_{i=t-j-k+1}^{t-j}\pi_i\delta_{t-i}\beta_{t-i})^2\right)\right), \tag{4.23}$$

and the residual $e_t(\boldsymbol{\delta}_{j,k})$ is given by

$$e_t(\boldsymbol{\delta}_{j,k}) = \begin{cases} e_t(\mathbf{0}) + \sum_{i=0}^{t-j}\pi_i\delta_{t-i}\beta_{t-i} & \text{if } t = j, \ldots, j+k-1 \\ e_t(\mathbf{0}) + \sum_{i=t-j-k+1}^{t-j}\pi_i\delta_{t-i}\beta_{t-i} & \text{if } t > j+k-1, \end{cases}$$

where $\pi_0 = -1$ , $\pi_i = \phi_i$ for $i = 1, \ldots, p$ and $\pi_i = 0$ for $i < 0$ and $i > p$. Therefore, the equation 4.23 can be written as

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) \propto \exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_{j,k}} e_t(\boldsymbol{\delta}_{j,k})^2\right),$$

and by replacing in (4.22) we obtain the probability (3.17) for any configuration of the vector $\boldsymbol{\delta}_{j,k}$. □

*Proof of Theorem 2.* Let $\boldsymbol{\theta}_{\beta_{j,k}} = (\boldsymbol{\phi}, \boldsymbol{\delta}, \sigma_a^2, \alpha)'$. The conditional distribution of $\boldsymbol{\beta}_{j,k}$ given the sample and the other parameters is

$$P(\boldsymbol{\beta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\beta_{j,k}}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\beta_{j,k}}; \boldsymbol{\beta}_{j,k}) \cdot P(\boldsymbol{\beta}_{j,k}).$$

The likelihood function $f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\beta_{j,k}}; \boldsymbol{\beta}_{j,k})$ is obtained in the proof of Theorem 1 (see equation (4.23)) and it can be expressed as

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\beta_{j,k}}; \boldsymbol{\beta}_{j,k}) \propto \exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_{j,k}}(e_t(0) + \boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k}\boldsymbol{\beta}_{j,k})'(e_t(0) + \boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k}\boldsymbol{\beta}_{j,k})\right).$$

Therefore,

$$\begin{aligned}
P(\boldsymbol{\beta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\beta_{j,k}}) \propto\ & \exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_{j,k}}(e_t(0) + \boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k}\boldsymbol{\beta}_{j,k})'(e_t(0) + \boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k}\boldsymbol{\beta}_{j,k})\right) \times \\
& \times \exp\left(-\frac{1}{2\tau^2}(\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}_0)'(\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}_0)\right) \\
\propto\ & \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}'_{j,k}\left(\frac{1}{\sigma_a^2}\sum_{t=j}^{T_{j,k}}\boldsymbol{D}_{j,k}\boldsymbol{\Pi}_{t-j}\boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k} + \frac{1}{\tau^2}\boldsymbol{I}\right)\boldsymbol{\beta}_{j,k}\right.\right. \\
& \left.\left. -2\left(-\frac{1}{\sigma_a^2}\sum_{t=j}^{T_{j,k}} e_t(0)\boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k} + \frac{1}{\tau^2}\boldsymbol{\beta}'_0\right)\boldsymbol{\beta}_{j,k}\right)\right) \\
\propto\ & \exp\left(-\frac{1}{2}(\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}^*_{j,k})'\boldsymbol{\Omega}^{-1}_{j,k}(\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}^*_{j,k})\right),
\end{aligned}$$

where $\boldsymbol{\Omega}_{j,k}$ and $\boldsymbol{\beta}^*_{j,k}$ are defined in (3.19) and (3.20) respectively. □